IN THE UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF TEXAS
HOUSTON DIVISION

DWIGHT BAZILE, et al,	§	
	§	
Plaintiffs,	§	
	§	
v.	§	**CIVIL ACTION NO. H-08-2404**
	§	
CITY OF HOUSTON,	§	**HONORABLE LEE ROSENTHAL**
	§	
Defendant.	§	

AFFIDAVIT OF WINFRED ARTHUR, JR.

STATE OF TEXAS)
)
COUNTY OF BRAZOS)

WINFRED ARTHUR, JR., appeared in person before me, the notary public whose signature and seal are subscribed below, and who knows the affiant to be the person whose signature appears below, confirming his identification by his Texas driver's license, number 06841899 and deposed as follows:

> I am a Full Professor of Psychology and Management at Texas A&M University. I received my PhD and M.A. in industrial/organizational psychology in 1988 and 1985, respectively, from the University of Akron. I have over 20 years of practical experience in the areas of test development, selection, public safety testing, and training. I am a Fellow of the Society for Industrial and Organizational Psychology, the American Psychological Society, and the American Psychological Association.
>
> In terms of my specific relevant experience, I have over twenty years of experience working with consulting firms and with public and private sector organizations in a capacity as an independent consultant and as Vice President of Barrett & Associates, Inc. I have performed job analysis and worked on the development and validation of small and large scale assessment systems. Relatedly, I have worked on a large number of projects involving job analysis and selection with fire, police and other public safety forces, including both entry level and promotional positions for fire and police.

I have also taught (and continue to teach) both graduate and undergraduate courses in personnel selection, and testing and psychometrics. I have directed students' dissertations, theses, and applied projects (practica) in these and other topic domains. I have published peer-reviewed papers in professional journals on topics pertaining to testing, psychometrics, personnel selection, and public safety selection. Additionally, I am the author of two books, over 15 book chapters, and over 60 refereed journal publications.

Concerning the issue of generating a scoring key **before** the administration and operational use of tests, the standard and expected scientific and professional practice is best summarized by Aiken (1988) who states that "professional test designers do not wait until a test is constructed and administered before deciding what scoring procedure to use." (p. 63). This view is consonant with that expressed in the *Standards for educational and psychological testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the *Principles for the validation and use of personnel selection procedures* (Society for Industrial and Organizational Psychology, 2003).

So, for instance, in the description of the test development process, the *Standards* make reference to four sequential phases of which the third is the "development, field testing, evaluation, and selection of the items and **scoring guides and procedures**" [emphasis added], and the fourth is "assembly and evaluation of the test for operational use." (p. 37). Thus, it is quite clear that the expectation, as articulated in the *Standards*, is that the scoring guides and procedures (scoring keys) will be developed **before** the administration and operational use of the test.

Although the specific nature and form of scoring keys and procedures may vary as a function of the form of the test (e.g., performance tasks vs. paper-and-pencil multiple-choice test vs. extended constructed-response tests [such as essays]), "all types of items require some indication of how to score the [test taker's] responses" (p. 38, AERA, APA, & NCME, 1999) and this must be specified before the test is administered and put into operational use. So, for tests in which the candidate is asked to select **the** best or correct answer from a number of alternatives (e.g., the typical multiple-choice test), one alternative is designated, a priori, as the correct or best answer. This alternative (or option) is technically described as the keyed alternative.

There are other test formats, methods, or instances where the alternatives may be differentially weighted. For example, a test method that is frequently characterized by the use of differentially weighted responses is the situational judgment test. With this test format, when candidates are required to, for example, (a) select the best and worst alternative, or (b) rate the effectiveness of each alternative, it may be argued that not all incorrect alternatives or options are equally wrong. Consequently, in such an instance, the scoring key specifies the weight that is to be assigned to each alternative, which in essence, allows for a range of scores for each item (i.e., full credit, some magnitude of partial credit,

and no credit; Motowidlo, Dunnette, & Carter, 1990; Ployhart & MacKenzie, 2011). In summary, regardless of the test format, the scientific and professional expectation is that the scoring key and procedures be explicated before the administration and operational use of the test.

Relatedly, from an applied and practical perspective, it is obviously impossible to develop a test item without some (foreknowledge) of what the keyed response or the expected correct or best answer is (e.g., see Haladyna, 1999, p. 77). Hence, knowledge of the correct or best answer to an item is the fundamental basis of any item development effort—that is, the test designer has to know what the focal knowledge, skill, ability, or other characteristic is, and then develop the test item to elicit or measure it. Furthermore, standard item development guidelines (e.g., Haladyna, 1999) are all predicated on the premise that the test designer has a keyed alternative or response as part of the item writing and development process. Consequently, the development of a scoring key occurs **concurrently** with, and not after the development of the test items. Indeed, even in situations that may appear to be exceptions, such as panel interviews, general scoring rubrics are developed and available before conducting the interviews.

In addition to being consonant with sound scientific and professional practice, other advantages to having a scoring key before the administration of a test include objectivity and transparency, and associated perceptions of test fairness.

It is also worth noting that the existence of a scoring key or guide before the administration of a test does not preclude post-administration changes to the key as warranted—for instance, as would occur in instances where test items are successfully challenged in post-administration reviews. In addition, the test designer may also use post-administration item analysis to identify possible mistakes in the scoring key (e.g., instances where there is inadvertently, more than one correct answer), or particularly poor items that were not caught in the item review process. Nevertheless, the possibility of these occurrences does not preempt or preclude the scientific and professional expectation and requirement for the development of a scoring key before the test is administered.

In summary, the scientific and professional expectation is that a scoring key or guide be developed before the administration of the test. Indeed, as previously noted, it is difficult to envisage how this could be otherwise since the development of keyed answers or responses is an integral part of the test development process such that the development of a key occurs concurrently with, and not after the development of the test items and subsequently, the test.

In formulating the opinions expressed herein, I have consulted and cited the following references:

Aiken, L. R. (1988). *Psychological testing and assessment* (6th ed.). Boston, MA: Allyn and Bacon.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Montowidlo, S.J. Dunette, M.D. & Carter, G.W. (1990) An Alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Ployhard, R.E. & MacKenzie, W.I. Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA Handbook of industrial and organizational psychology: Volume 2, Selecting and developing members for the organization* (pp. 237-252). Washington, DC: APA.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

I declare the foregoing is my opinion on these matters.

Winfred Arthur Jr.
WINFRED ARTHUR, JR.

19 SWORN TO AND SUBSCRIBED BEFORE ME, the undersigned notary public on this day of August, 2010.

Lora Adams
Notary Public By and For The State of Texas

My Commission Expires:
July 18, 2012

Notary's Name Printed or Typed:
Lora Adams

LORA ADAMS
Notary Public, State of Texas
My Commission Expires
JULY 18, 2012